UNIVERSITY HENRI POINCARE, NANCY

LCM3B

(Laboratoire de Cristallographie et Modélisation de Matériaux Minéraux et Biologiques)

ECM 19

First announcement

DIRECT PHASING IN CRYSTALLOGRAPHY :

STATISTICAL APPROACH WITH MULTIMINIMA SCORE FUNCTIONS

IMPB, Pushchino, Russia :

V. Lunin

N. Lunina

T. Petrova

T. Skovoroda

E. Vernoslova

IGBMC, Strasbourg, France

A. Podjarny

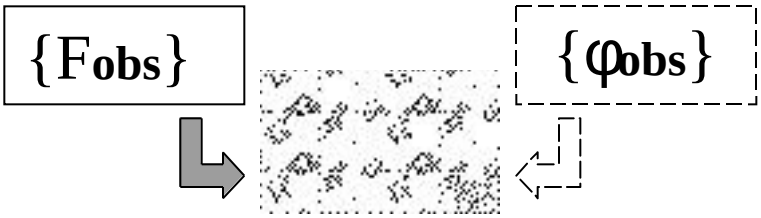
LCM3B, University of Nancy, France

E. Chabrière

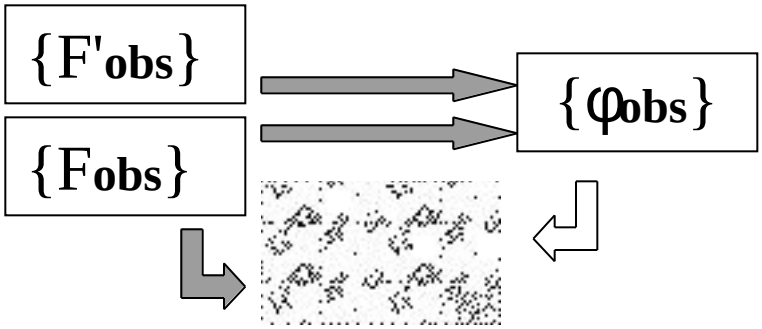
A. Urzhumtsev

$$\rho(\mathbf{r}) = \Sigma_{\mathbf{s}}\, F(\mathbf{s}) \exp\{i\varphi(\mathbf{s})\} \exp(-2\pi i \mathbf{sr})$$

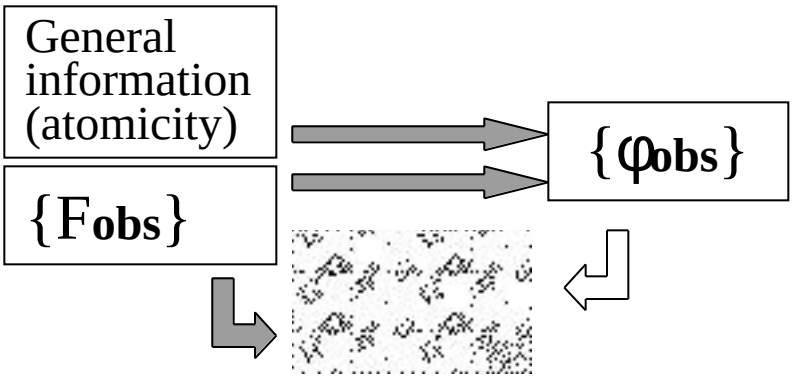
PHASE PROBLEM



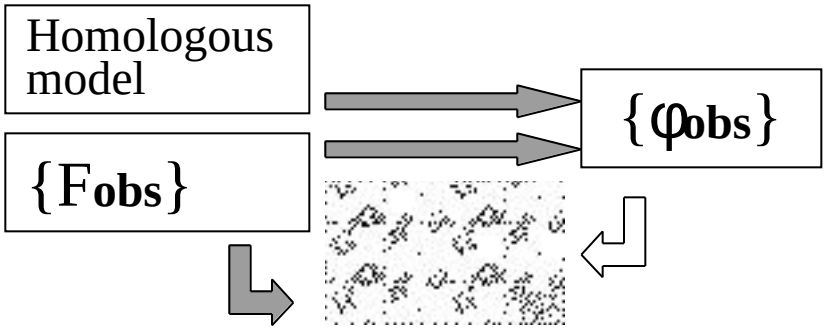
'**Experimental phasing'** by writing and solution of some equations (MIR, MAD)



Direct methods ; Initially - solution of some equations; minimisation approach by Sayre & Toupin, 1975



Molecular replacement ; Minimisation approach



PHASE PROBLEM :
MINIMISATION APPROACH

2. How to minimise ?

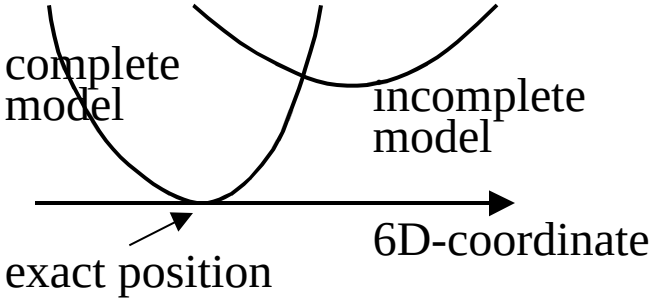
1. To which model (to which phase set) corresponds the point of minimum ?

0. Which information and which search (selection) to use ?

Lunin, Urzhumtsev, Skovoroda (1990) *Acta Cryst*., **A46**, 540

Baker, Krukowski, Agard (1993) *Acta Cryst*., **D49**, 186

Example : Molecular replacement

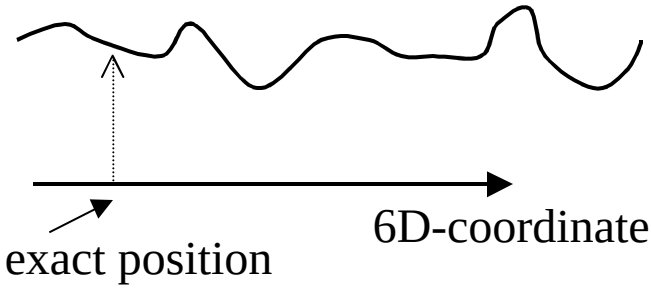


Ideal case :

no experimental errors, correct model

Real case:

incomplete model with errors



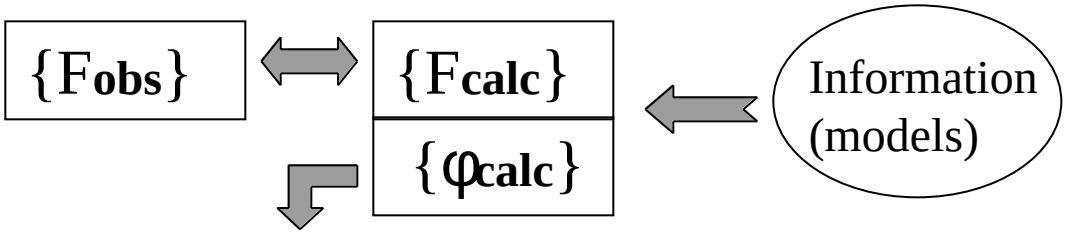
DIRECT SEARCH OF THE SOLUTION

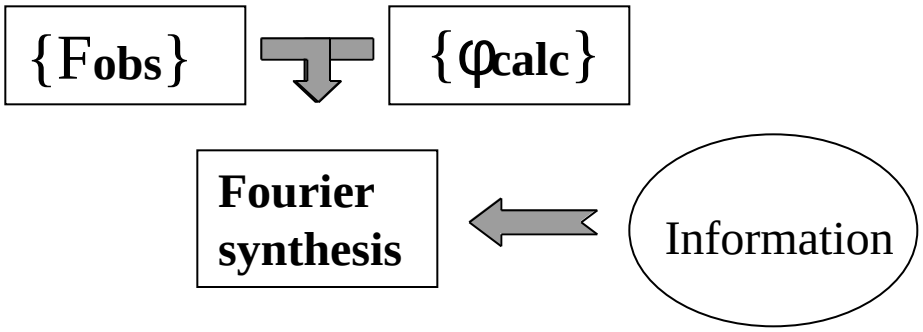
A single set of {Fobs} + general information

Two strategies :

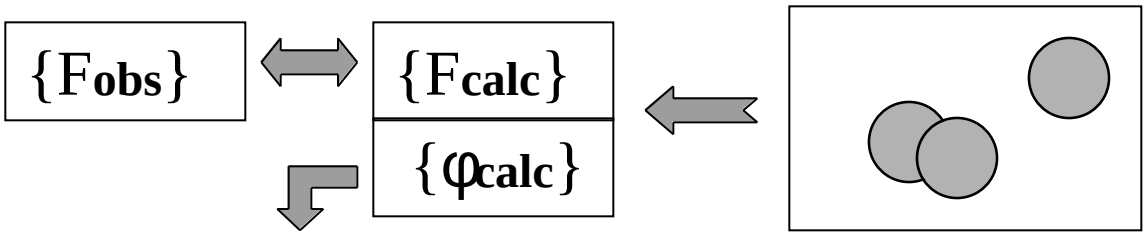
- use the whole data set (more information)
- use few data at the beginning, for example starting from the low-resolution end (smaller search space; search functions eventually behave better)

Use of general information for the phase selection, two ways :





SEARCH WITH MODELS

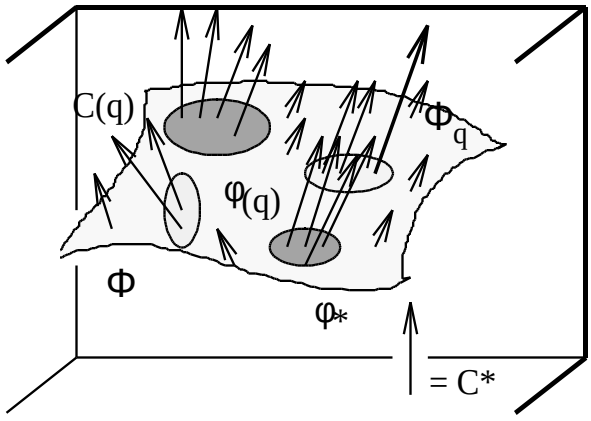


$$C = Corr(F_{calc}, F_{obs}) \_ \rightarrow max\ ?$$

Exemple : Few Atoms Model method :

Lunin et al. (1995) *Acta Cryst.*, **D51,** 896-903

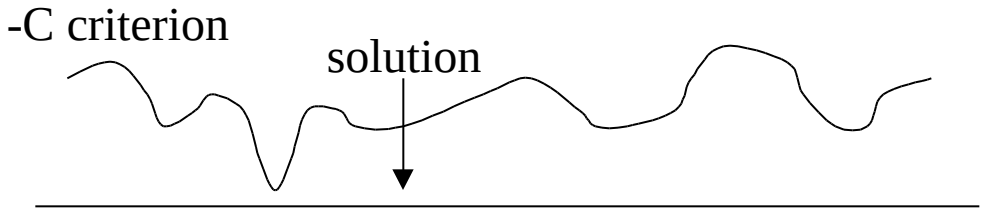
+ faster search
- selection is indirect
- quality of models

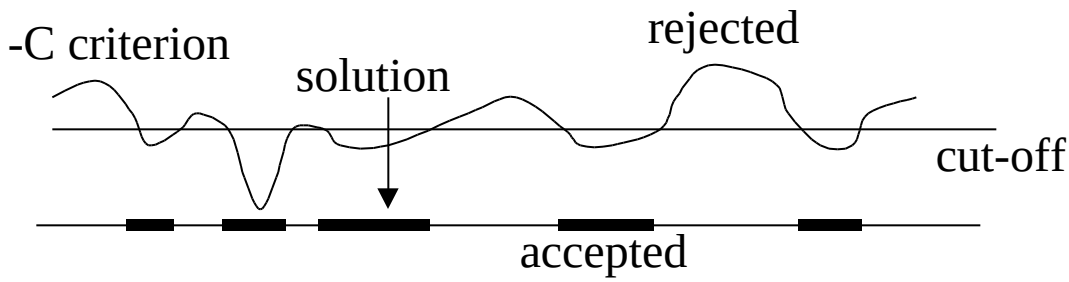


The problem is neither in the quality of the models nor in the least-squares or correlation criterion.

If calculated structure factors magnitudes are close to the experimental ones, the corresponding phase set in not necessarily close to the correct solution.

USUAL CASE OF SEARCH CRITERION

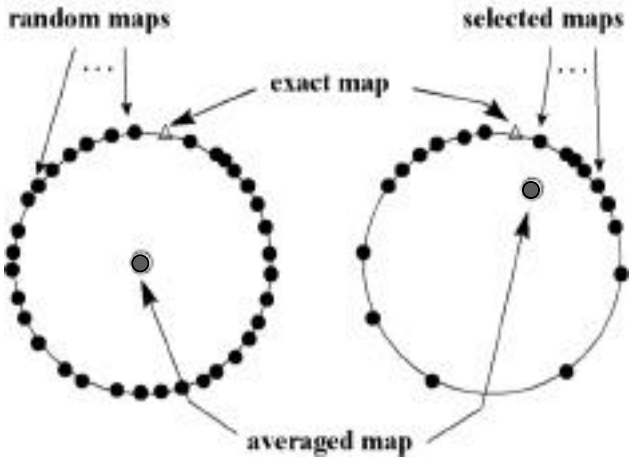




R becomes a binary (selection) criterion

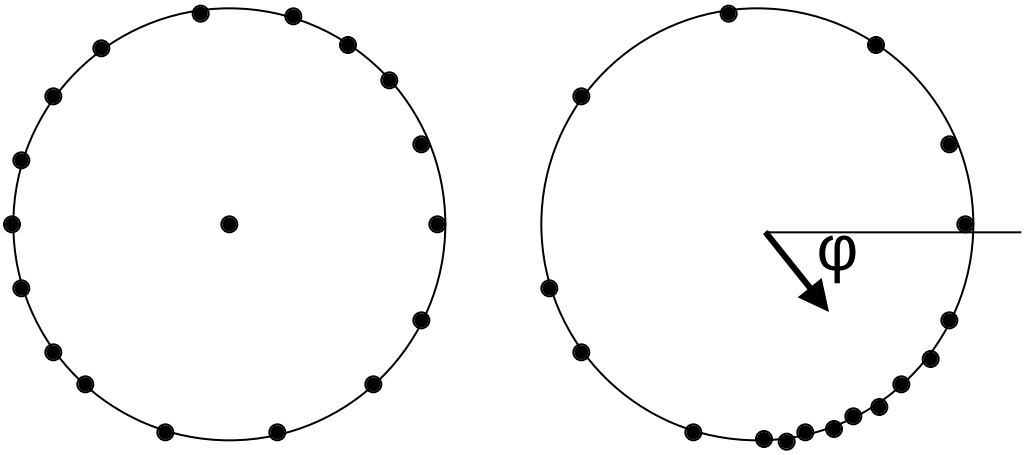
Any INDIVIDUAL PHASE SET can be relatively correct or completely wrong.

However, the MEAN VALUE OF THE SELECTED PHASES is better that the mean value of randomly generated phases.



Clustering analysis can be used to improve the result

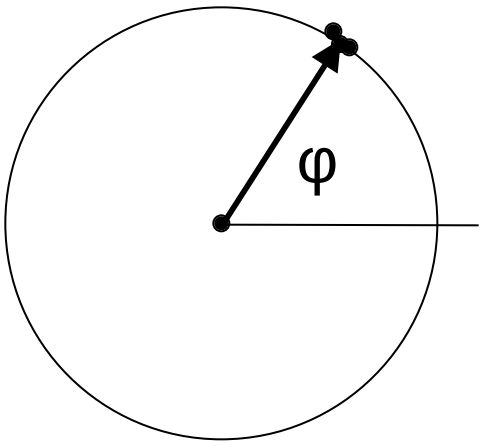
PHASE AVERAGING
(AFTER ALIGNMENT)



m = 0 m = 0.5

For any reflection corresponding phases from all K selected phase sets are averaged

$$m \exp(i\varphi) = \Sigma_k \exp(i\varphi_k) / K$$

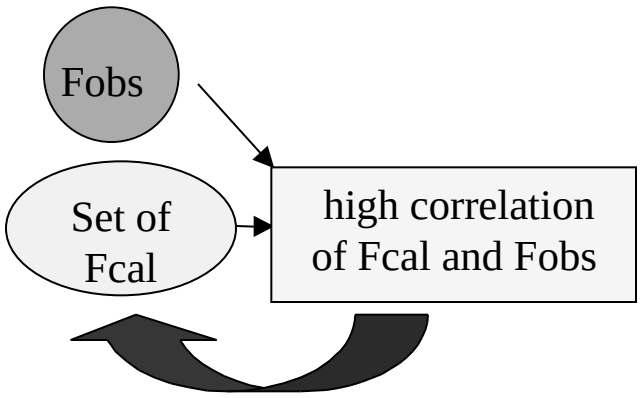


m = 1

STRUCTURE FACTORS CONSTRAINTS

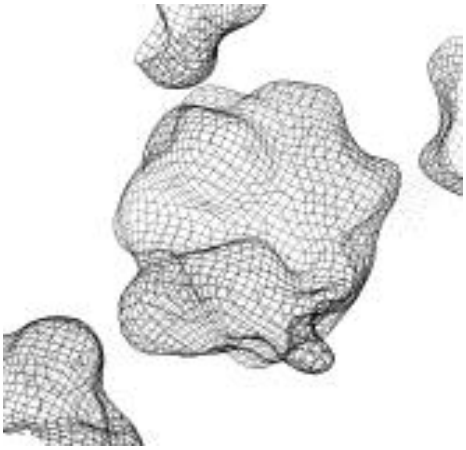
Example : FAM method (*Lunin et al., 1995, 1998*).

If a calculated structure factors magnitudes are close to the experimental ones, is the corresponding phase set close to the correct solution ?



Is the phase set good ?
Not necessarily

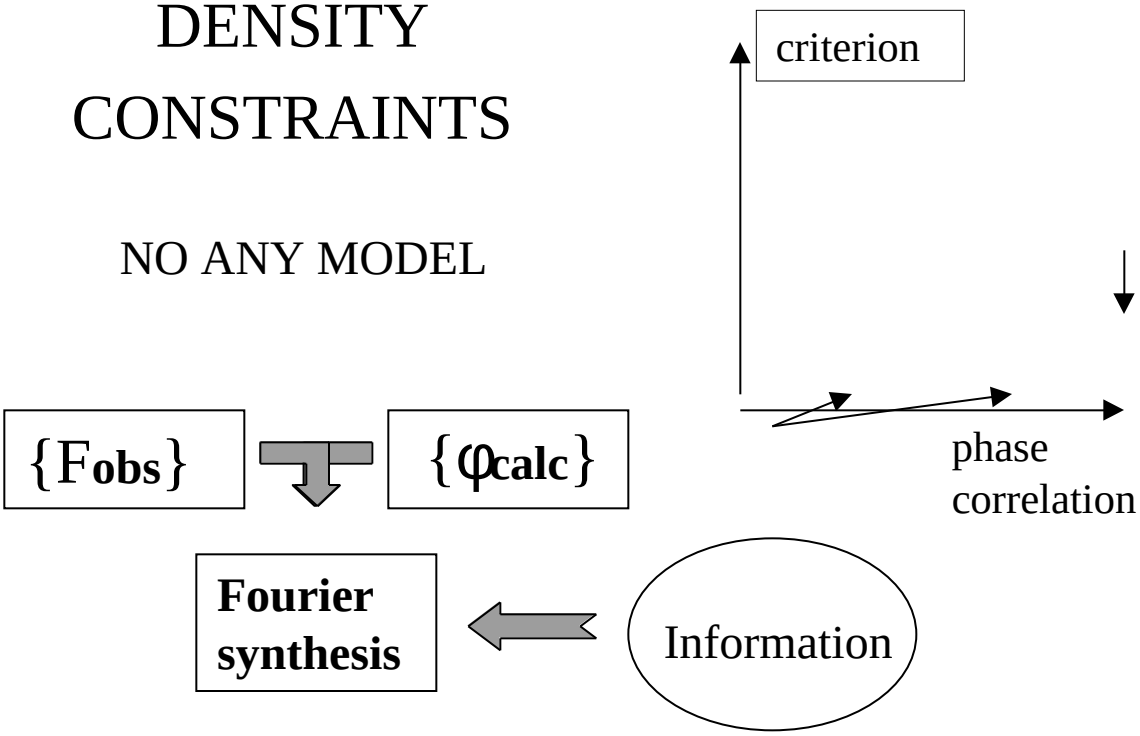
Example of a practical application *(Urzhumtsev et al., 1996; Lunin et al., unpublished):*



Molecular envelope for the 50S particle from *Thermus thermophilus* found *ab initio* by the FAM method, experimental data by A.Yonath.

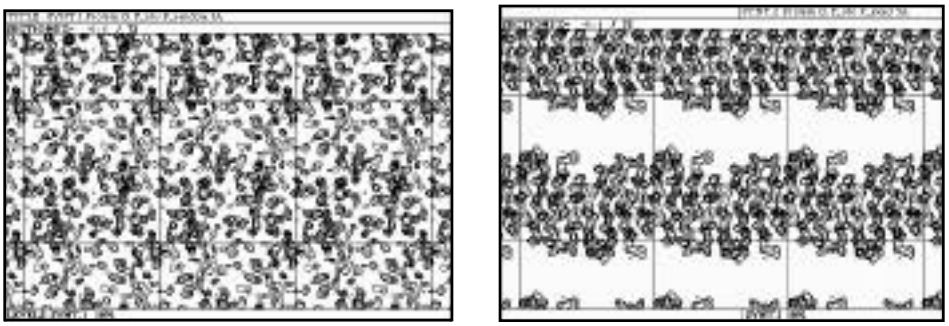
DENSITY CONSTRAINTS

NO ANY MODEL



Can we distinguish between a ‘bad’ and a ‘good’ maps ?

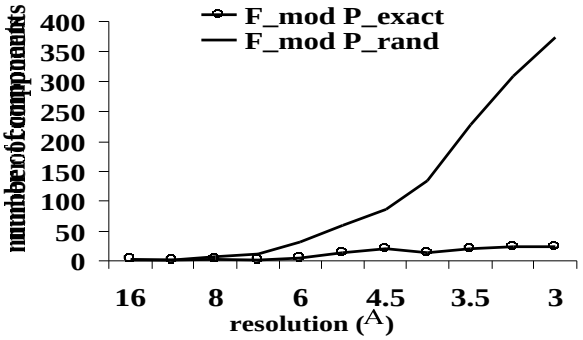
Protein G; 3Å resolution (1323 reflections)

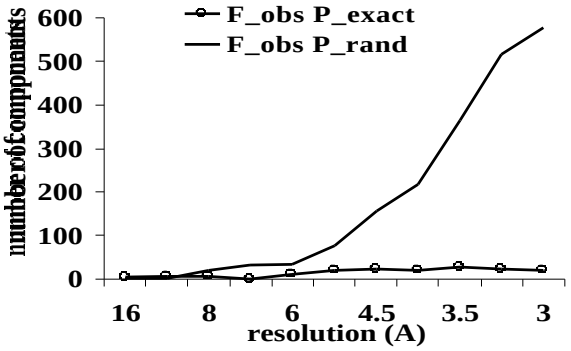


F exact, φ random F exact, φ exact

An information, important for map interpretation, is rather the number, position and shape of connected regions and peaks of Fourier synthesis and than the exact value of the density

CONNECTIVITY





Protein G

10% of the unit cell volume are analysed (25 Å^3 per residue)

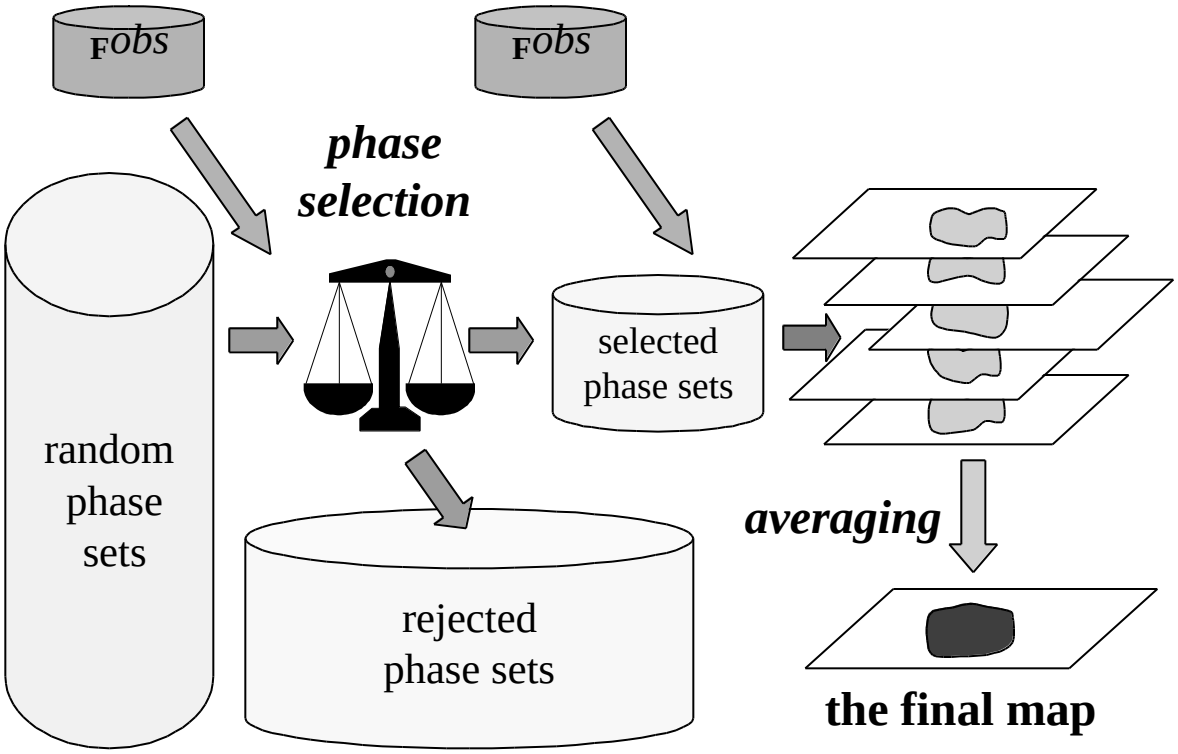
F_mod, P_exact are calculated from the refined atomic model

Basis

It is expected to see several compact globular regions at the correct low-resolution synthesis. We look for phases which being coupled with experimental magnitudes result in Fourier synthesis revealing the number of globular regions, each of the same finite volume, equal to the number N_{mol} of molecules in the unit cell.

Other topological criteria are possible

RANDOM CONNECTIVITY-BASED SEARCH



Number M of generated random phase sets $\{\varphi_s\}_m$: very large.

Every $\{\varphi_s\}_m$ is used together with Fobs to calculate a map.

Search goal: the phase set is **selected** if the corresponding map has desired connectivity properties (chosen cut-off level κ^*):

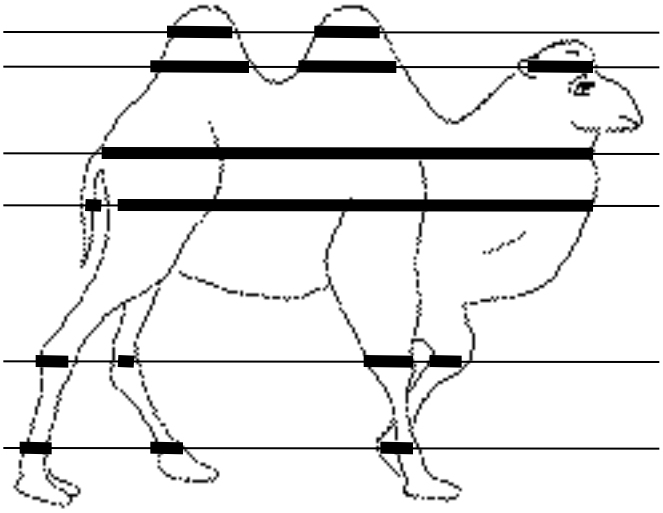
For example, that the Ω_κ region consists of *Nmol* equal (or practically equal) connected components where *Nmol* is the number of macromolecules in the unit cell

The selected phase sets are **aligned** and averaged in order to get the 'best' phases and corresponding figures of merit :

- $\rho_1(\mathbf{r})$ and $\rho_2(\mathbf{r})$ calculated with Fobs and $\{\varphi_s\}_1$ or $\{\varphi_s\}_2$;

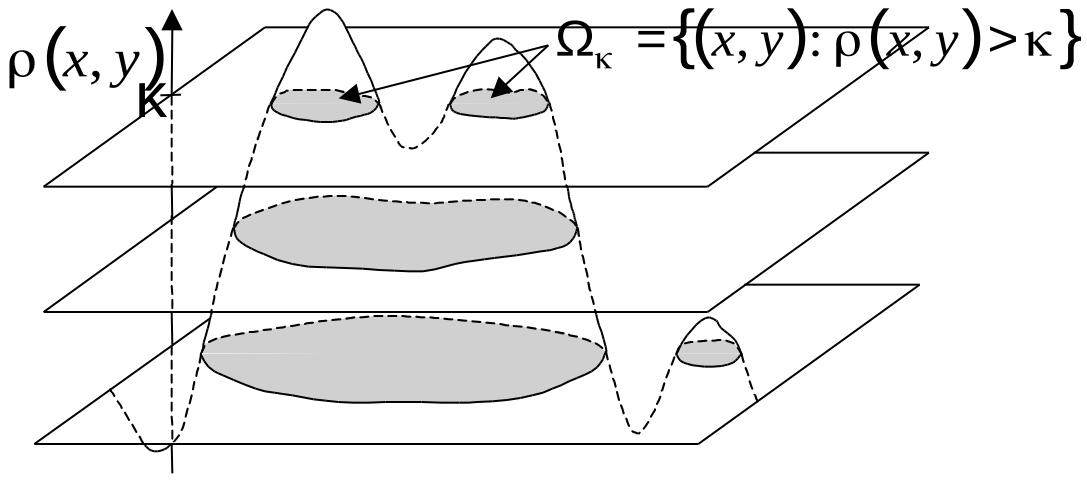
admissible origin shifts (+ enantiomer choice if possible) are applied in order to find the highest possible map correlation C_φ.

CONNECTIVITY



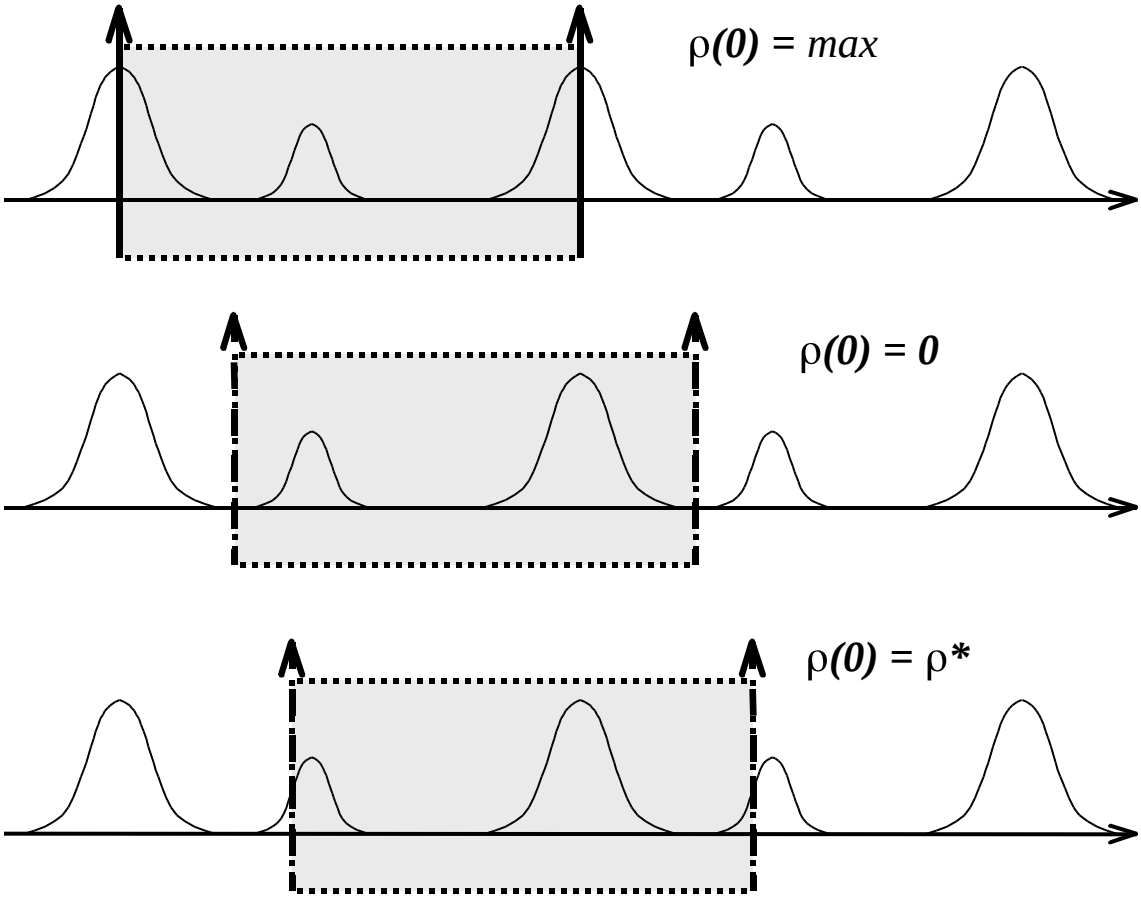
For different cut-off levels κ the regions $\Omega_\kappa=\{\mathbf{r} : \rho(\mathbf{r})>\kappa\}$ in the unit cell are analysed: the number of connected components and their volume are determined.

Example: Two-dimensional connectivity analysis



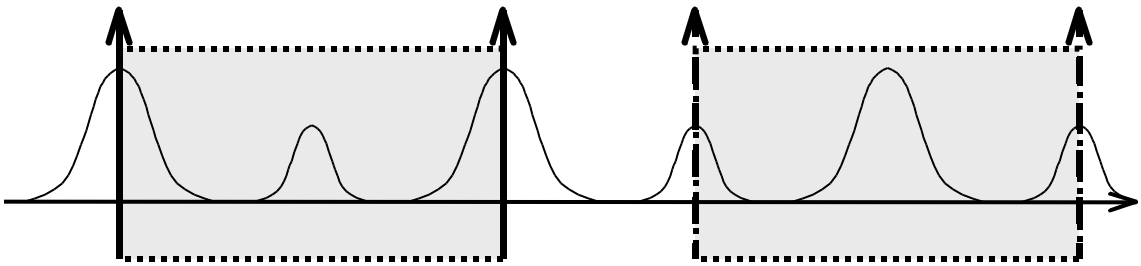
ORIGIN CHOICE OF THE UNIT CELL

General case



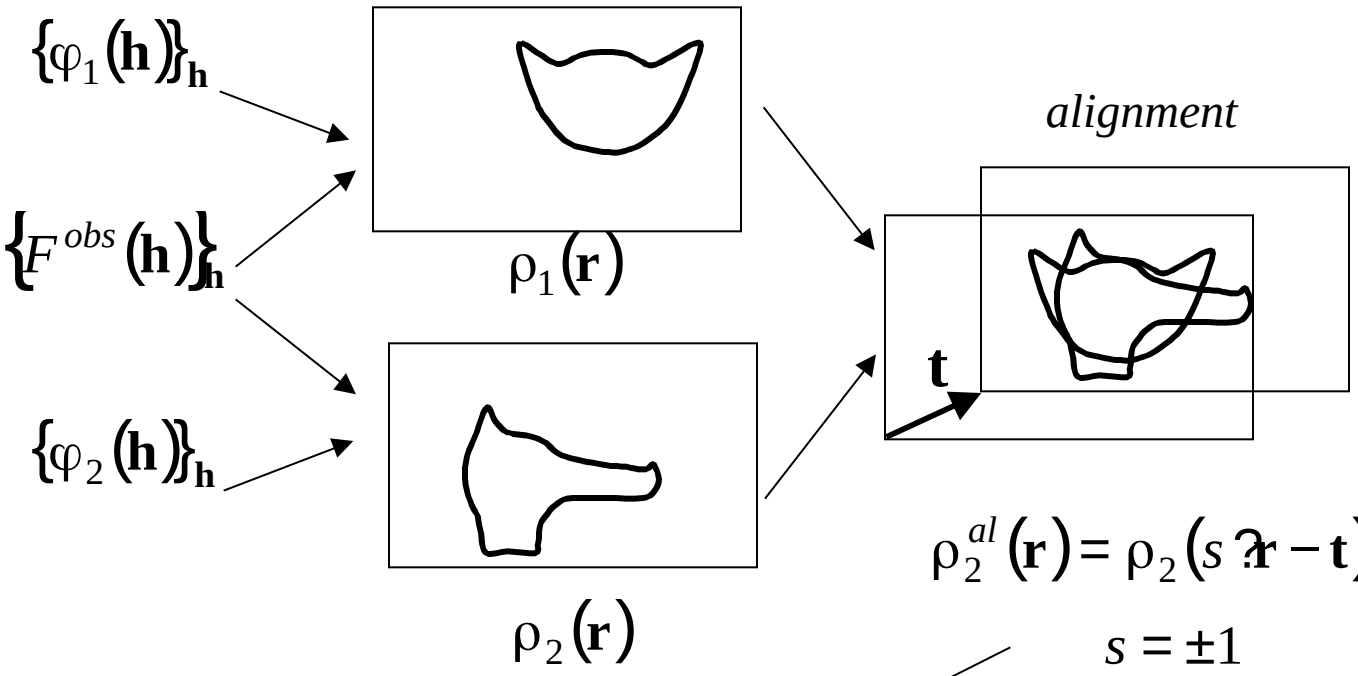
ρ***(r)*** = ρ***(-r)***

Particular symmetry



ALIGNEMENT OF PHASE SETS (MAPS)

Lunin & Lunina (1996) *Acta Cryst.*, **A52**, 365-368

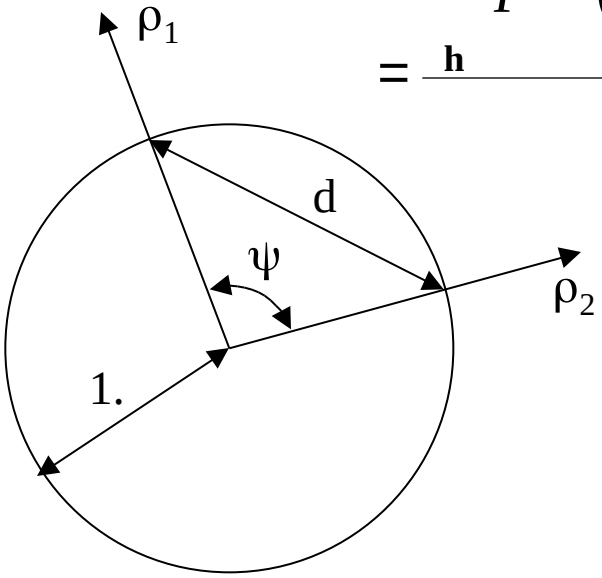


$$\rho_2^{al}(\mathbf{r}) = \rho_2(s\,\mathbf{r} - \mathbf{t})$$

$$s = \pm 1$$

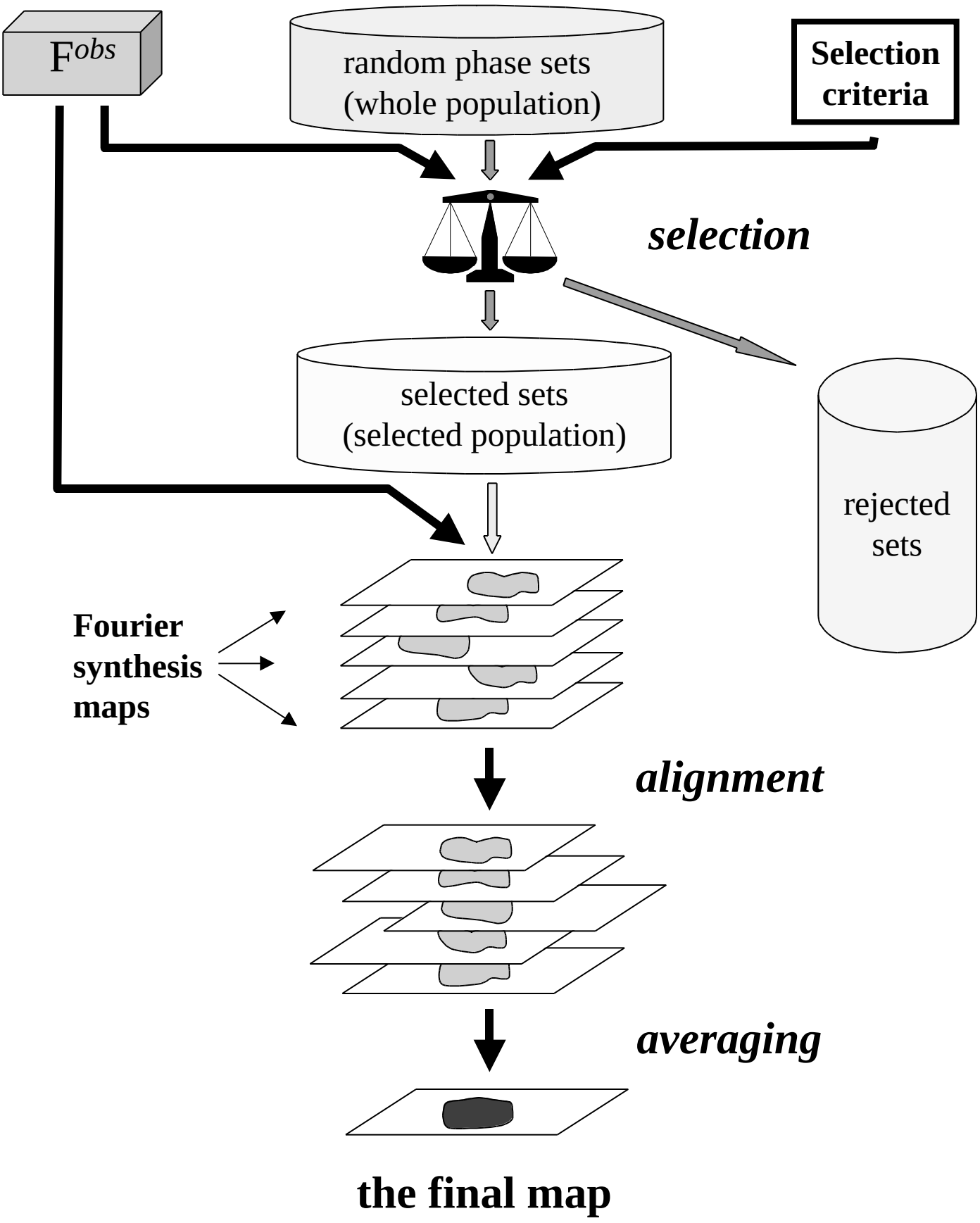
$$C_\varphi(\rho_1, \rho_2) = \frac{\rho_1(\mathbf{r})\,\rho_2^{al}(\mathbf{r})\,dV_\mathbf{r}}{\sqrt{\rho_1(\mathbf{r})^2\,dV_\mathbf{r}\;\;\rho_2^{al}(\mathbf{r})^2\,dV_\mathbf{r}}}$$

$$= \frac{\sum_\mathbf{h} F^{obs}(\mathbf{h})^2 \cos[\varphi_1(\mathbf{h}) - s\,\varphi_2(\mathbf{h}) - 2\pi(\mathbf{h},\mathbf{t})]}{\sum_\mathbf{h} F^{obs}(\mathbf{h})^2}$$



$$C_\varphi(\rho_1, \rho_2) = \cos\psi$$

$$\mathrm{d} = dist(\rho_1, \rho_2) = \sqrt{2(1. - C_\varphi)}$$


F^obs
random phase sets
(whole population)
Selection
criteria
selection
selected sets
(selected population)
rejected
sets
Fourier
synthesis
maps
alignment
averaging
the final map

TEST OBJECT 1. γ-CRISTALLIN IIIb

$P2_12_12_1$ (N_{sym}=4);
a=58.7Å, b=69.5Å, c=116.9Å;
2 molecules / asymmetric unit,
173 residues each.
The structure solved by:
Y. Chirgadze *et al.* (1991)
Exper. Eye Res., 53, 295-304.
28 reflections (24Å resolution)
were used for *ab initio* phasing

F^{obs} - experimental; φ^{exact} - calculated from the atomic model were used to estimate the results of the phasing.

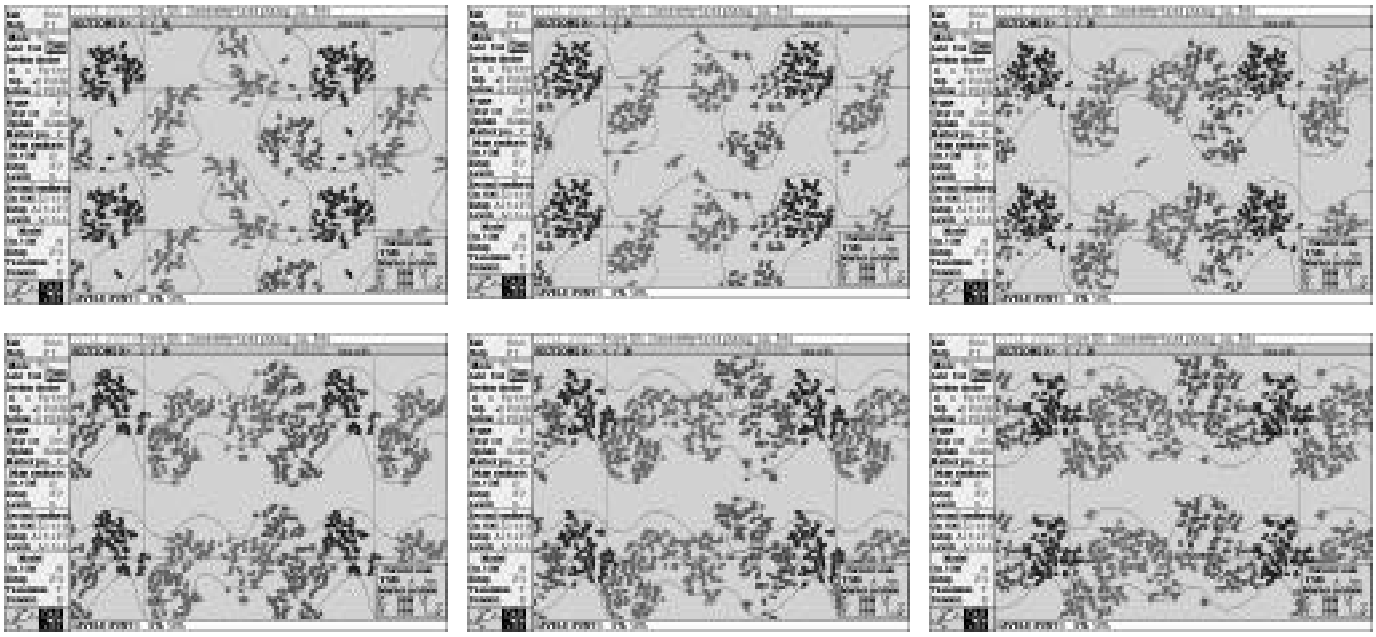
Selection criterion: Ω_κ consists of 8 finite connected components. The two components connected by n/c symmetry are different in volume not more than 10%.

Ω_κ volume per residue (in Å ^{3})	number of connected components and their sizes (in grid points)
5.	4*303+4*279
15.	4*856+4*835
25.	4*1402+4*1398
30.	2*6676
35.	1*15640

TEST RESULTS: γ-CRISTALLIN IIIb

The comparison of the exact phase set with 100 000 randomly generated and 495 connectivity-selected phase sets shows that
- there are quite wrong phase sets which nevertheless result in maps possessing desired connectivity properties;
- .5the concentration of good phase sets is significantly higher in the selected sets than in the random population.

The alignment and averaging of the maps corresponding to the selected phase sets allows to get the map with the correlation (with the exact one) equal to 0.89 for 28 reflections included.



All sections of the average map are shown for the independent part of the unit cell superposed with the atomic positions for the refined model. The levels correspond to the relative volume of Ω_K equal to 10 (black) and 200 (blue) $Å^3$/res. Differently coloured atomic positions correspond to symmetry related molecules.

TEST OBJECT 2. RNAse *sa*

$P2_12_12_1$ (N_{sym}=4);
a=64.9Å, b=78.32Å, c=38.79Å;
2 molecules / asymmetric unit,
96 residues each;
The structure solved by:
_ev_ik *et al*. (1991)
Acta Cryst., **B**47, 240-253.

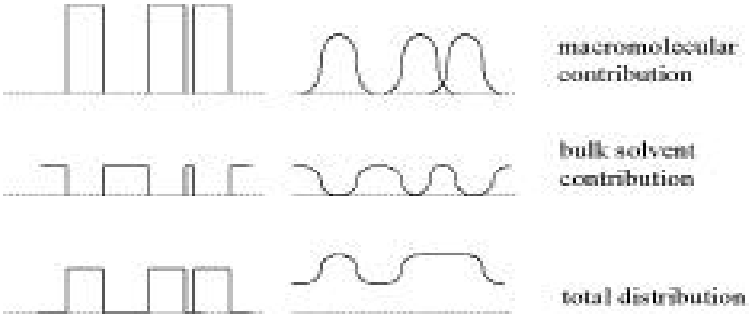
29 reflections (18Å resolution) were used for *ab initio* phasing

Selection criterion: Ω_κ consists of 8 finite connected components. The two components connected by n/c symmetry are different in volume not more then 10%.

Ω_κ **volume** per residue (in Å ^{3})	number of connected components and their size (in grid points)
5.	4*101+4*24+4*3
15.	2*570+4*91
25.	2*1246
30.	2*1734
35.	1*2238

In this case the exact phase set does not satisfy the selection condition due to too close packing and the contribution of the solvent

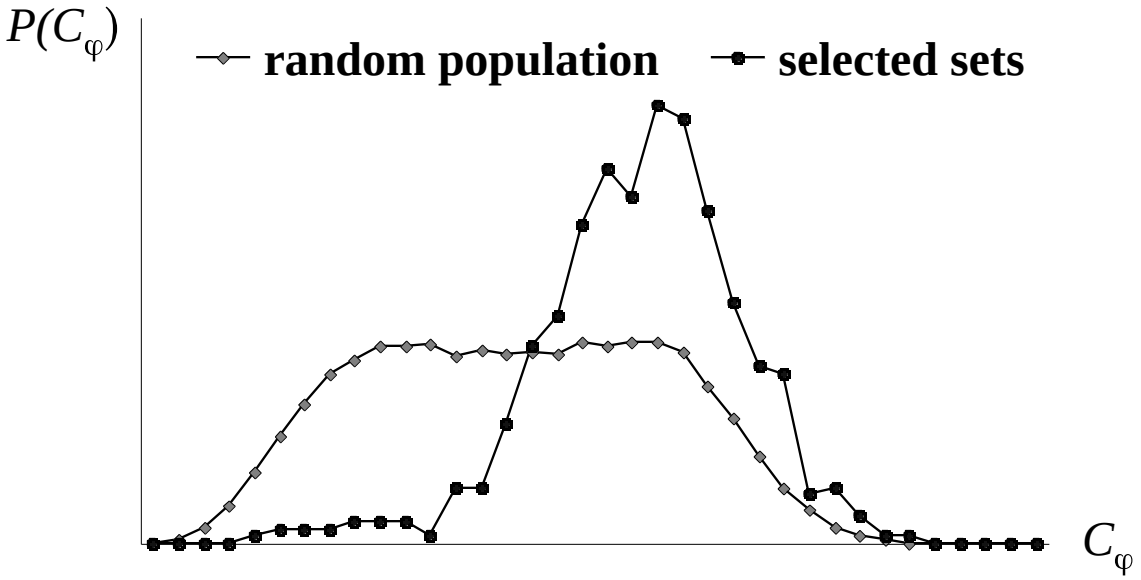
Nevertheless, let's try to search for 8 domains



ideal finite resolution

RESULTS OF THE SEARCH

The comparison of the exact phase set with 100 000 randomly generated and 495 connectivity-selected phase sets shows that:
- there are quite wrong phase sets which nevertheless result in maps possessing desired connectivity properties;
- the concentration of good phase sets is significantly higher in the selected sets than in the random population.

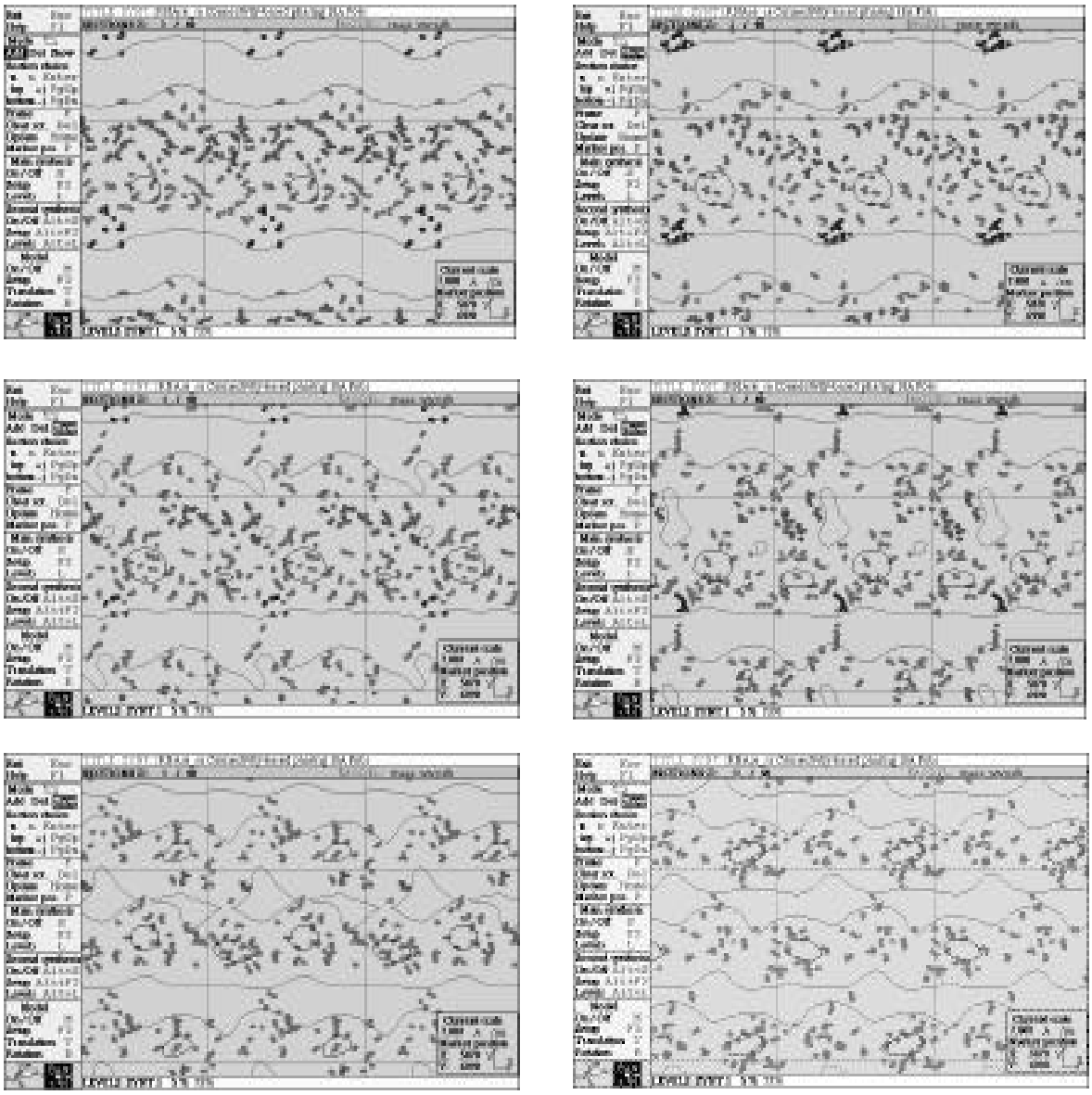


Distribution of the value of the correlation C_φ of the exact 18Å map with the one calculated with the observed magnitudes and trial phases

The alignment and averaging of the maps corresponding to the selected phase sets allows to get the map with the correlation (with the exact one) equal to 0.72 if all 29 reflections are included.

For the 24Å resolution map the correlation with the exact map is 0.91 (13 reflections).

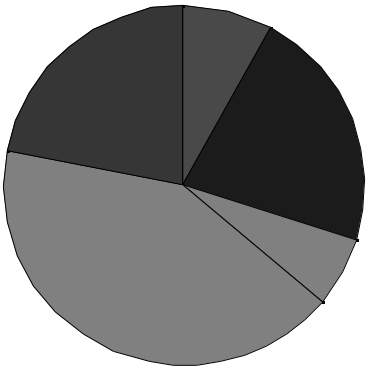
RESULTS OF THE SEARCH : RNAse



All sections of the average map are shown for the independent part of the unit cell superposed with the atomic positions for the refined model. The levels correspond to the relative volume of Ω_K equal to 10 (black) and 200 (blue) $Å^3$/res. Differently coloured atomic positions correspond to symmetry related molecules.

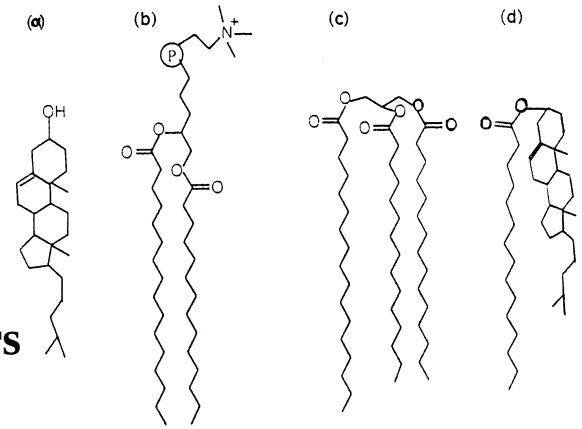
LDL : LOW DENSITY LIPOPROTEIN

Human Low Density Lipoproteins are the major cholesterol carriers in the blood. Elevated concentration of LDL (“bad cholesterol”) is a major risk factor for atherosclerotic disease, coronary heart disease, insulin resistance syndrome.



- **cholesterol**
- **phospholipids**
- **triglycerides**
- **cholesterol esters**
- **ApoB protein**

(550Kda, 4536 aa)



Electron microscopy studies : Luzzati, Tardieu & Aggerbeck (1979), van Antwerpen et al. (1997), Orlova et al. (1999).

LDL particle : ~ 220Å in diameter (180-250Å for different fractions)

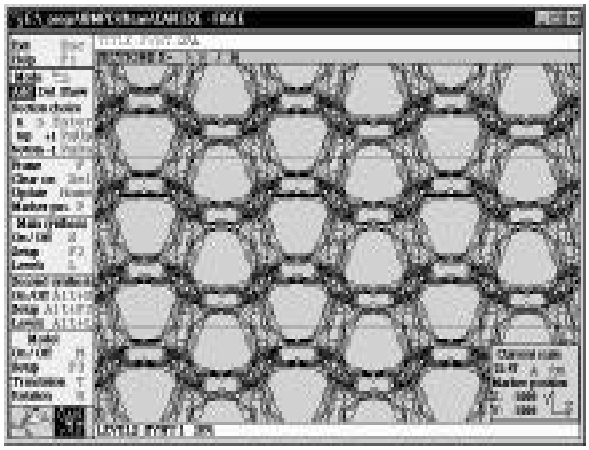
Crystallisation : Prassl et al. (1996), Ritter et al. (1997).

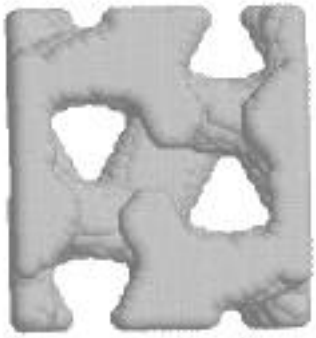
Space group : C2, unit cell is 180*416*379 Å,

27 Å resolution data set, very complete (about 800 reflections).

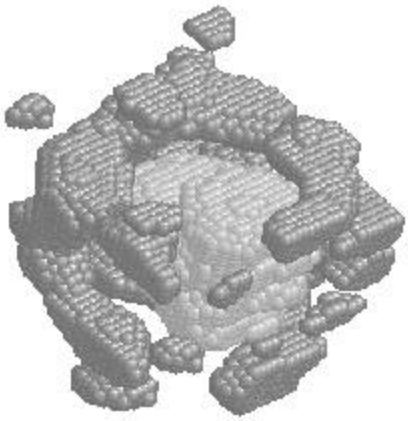
LDL: CONNECTIVITY-BASED SEARCH

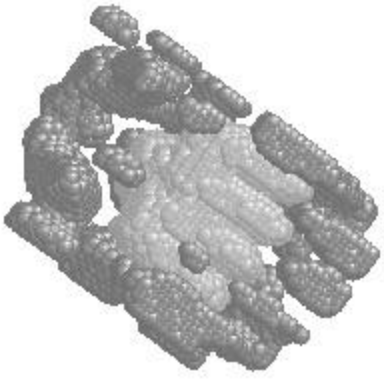
Experimental data : Ritter,S., Diederichs, K., Frey I., Berg, A., Keul, J., & Baumstark, M. (1999) J. of Crystal Growth, **196**, 344-349 (Universities of Freiburg and of Konstanz, Germany)





The isolated connected regions, identified by the connectivity search, are the LOW-density ones corresponding for the lipid core. The particles are packed very densely, forming very close surface contacts of the apoB.

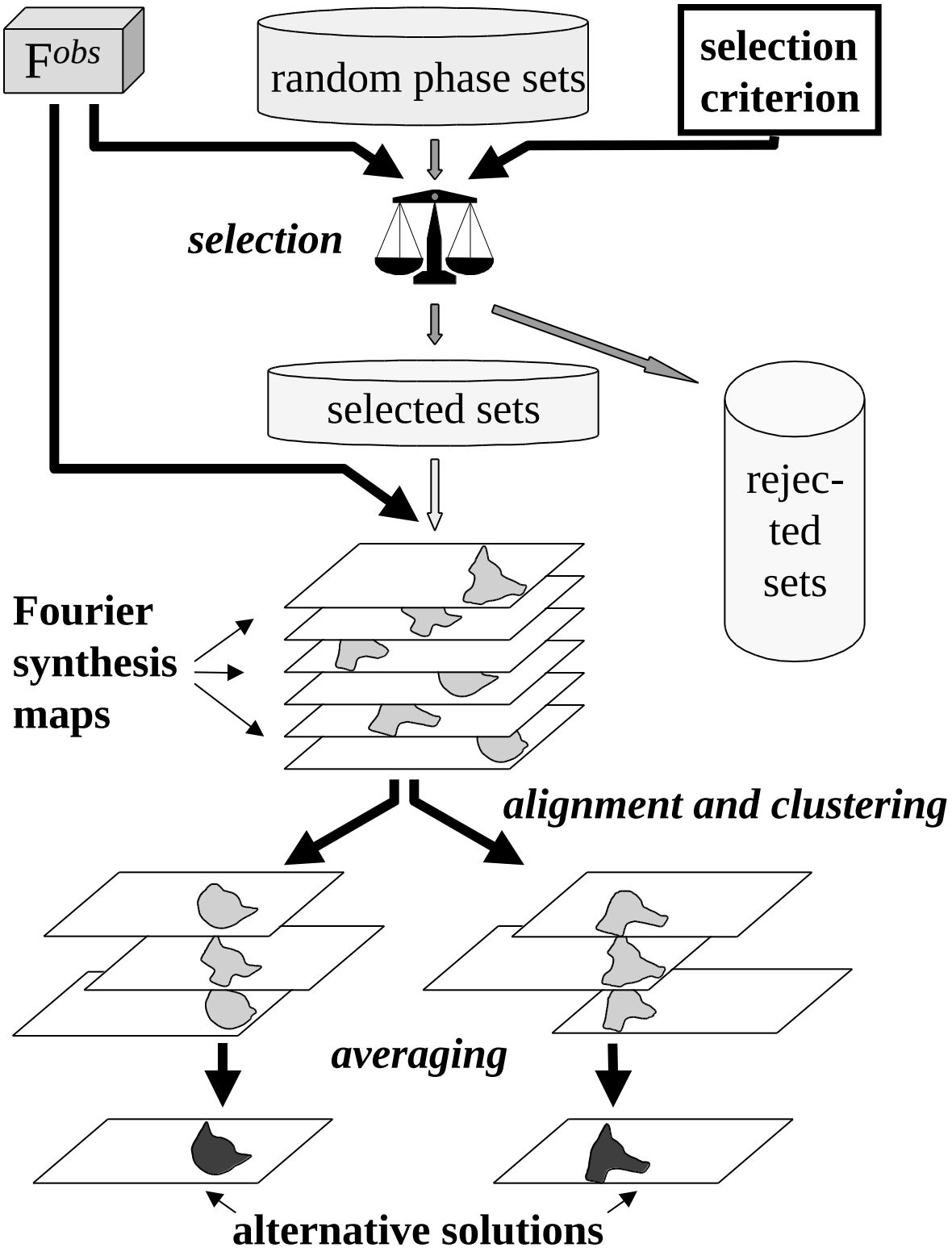




Lunin, V.Yu., Lunina, N.L., Ritter, S., Frey, I., Keul, J., Diederichs, K., Podjarny, A., Urzhumtsev, A.G., Baumstark, M. (2001) "Low-Resolution Data Analysis for the Low-Density Lipoprotein Particle". *Acta Cryst.*, **D57,** 108-121

GENERAL SCHEME OF LR DIRECT PHASING

(Lunin, Urzhumtsev, Skovoroda (1990) *Acta Cryst.*, **A46**, 540-544)



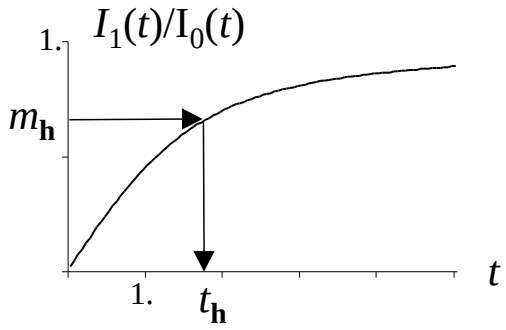
PHASE EXTENSION : GENERATION MODES

1. ***No preliminary phase information***

The phase ϕ_h : uniform distribution

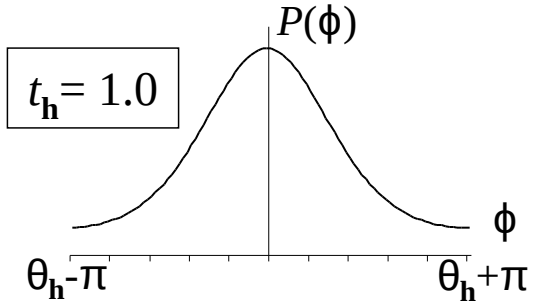
- *acentric reflections*: in [0,2π] interval ;
- *centric reflections*: as ϕ^0 or $\phi^0+\pi$ (ϕ^0 is an allowed phase value).

2. ***An approximate phase value*** θ_h ***and its figure of merit*** m_h ($0 \le m_h \le 1$) ***are known***



The phase ϕ_h is generated in accordance with the distribution

$$P(\varphi) \quad \exp\left[t_h \cos(\varphi - \theta_h)\right]$$



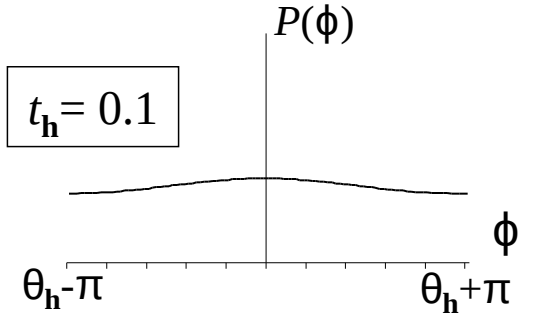
reliable phase, large m_h, t_h

The distribution parameter t_h is chosen to satisfy the condition

$$\langle \cos(\varphi - \theta_h) \rangle = m_h$$

which is

$$I_1(t_h)/I_0(t_h) = m_h$$

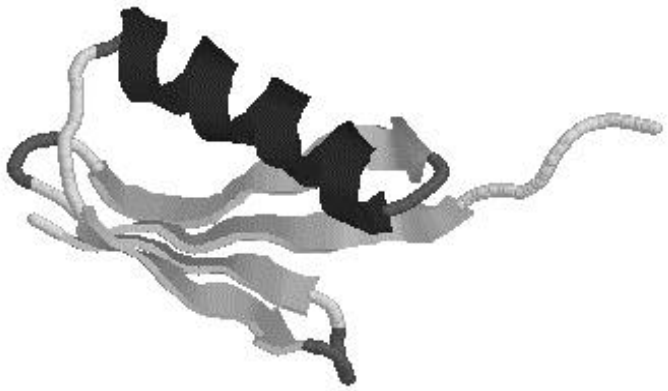


unreliable phase, small m_h, t_h

PHASE EXTENSION : PROTEIN G

1IGD
Protein G
Immunoglobulin Binding Protein
Derrick, J.P. & Wigley, D.B. (1994) *J. Mol. Biol.*, **243**, 906.

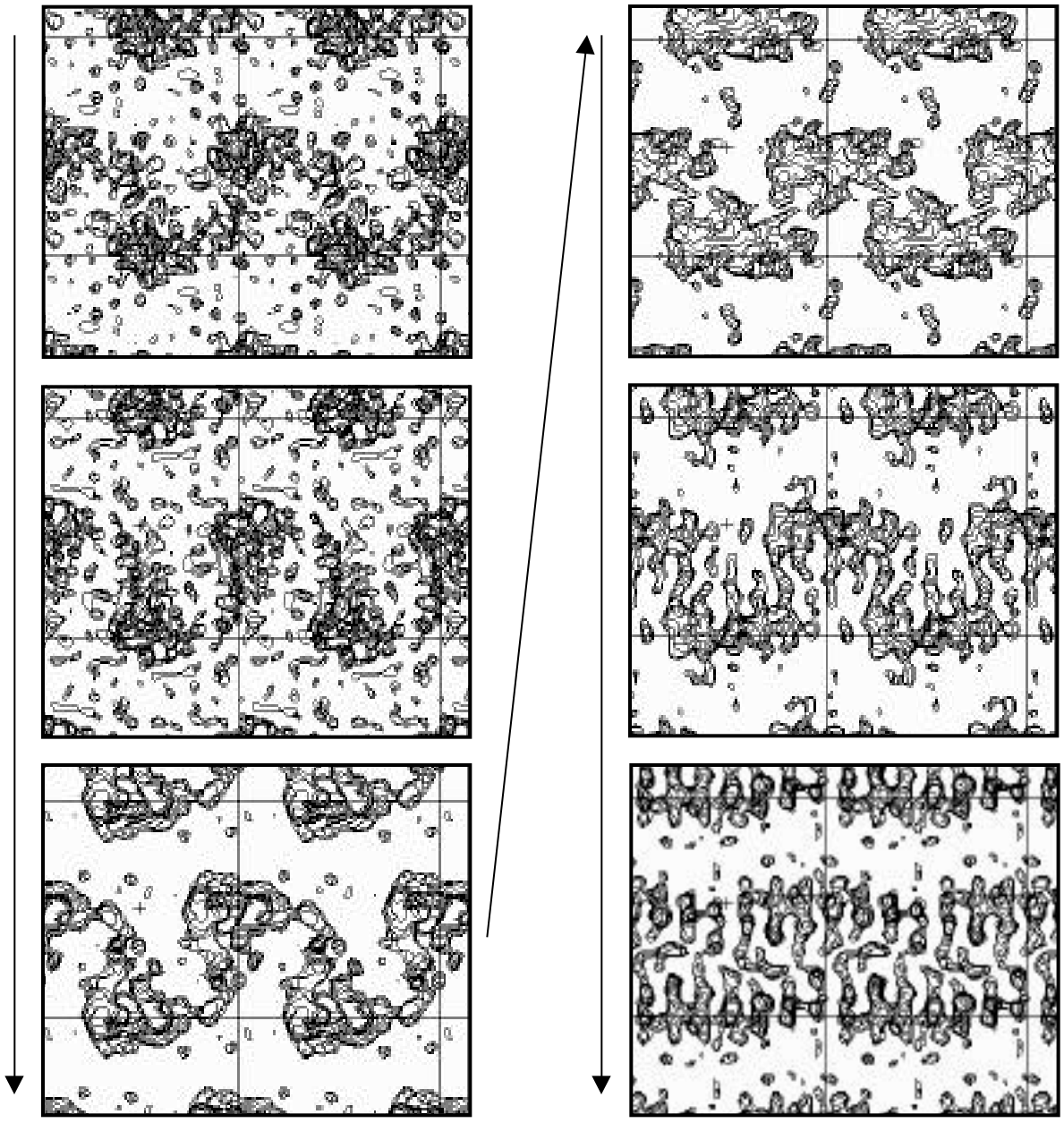
$P2_12_12_1$;
34.9 40.3 42.2 (Å);
90. 90. 90. (°)
61 residues;
588 atoms.



Number of reflections :

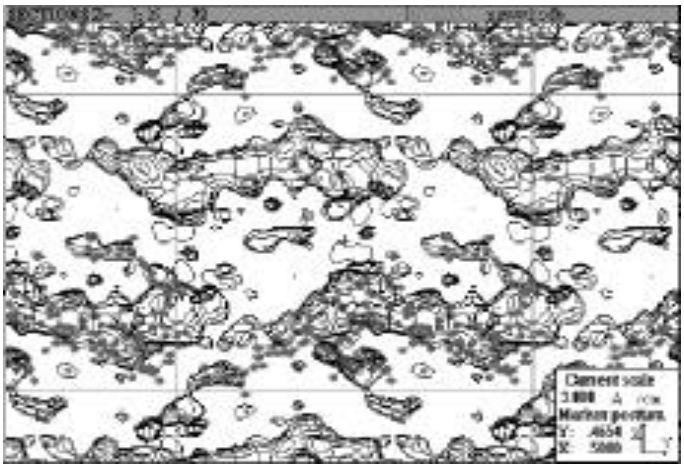
∞-16	∞-12	∞- 8	∞- 6	∞- 5	∞- 4	∞- 3.5	∞- 3
15	28	85	181	305	580	847	1323

PROGRESS OF THE PHASE EXTENSION

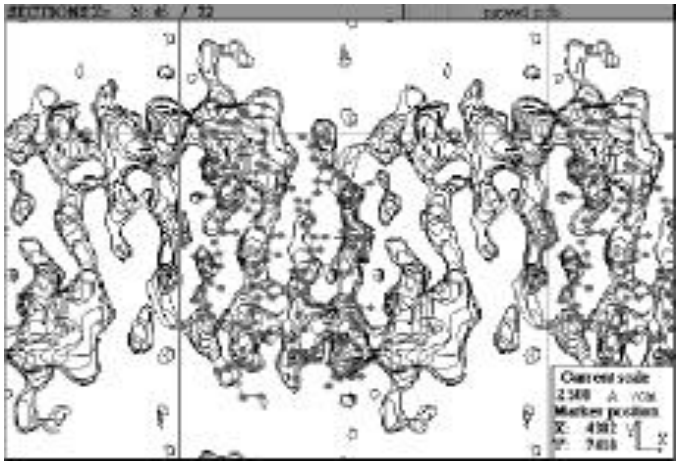


projections along z-axes of a part -6/72 < z < 6/72 of weighted 3Å Fourier syntheses at different stages of phasing are shown

SECONDARY STRUCTURE ELEMENTS



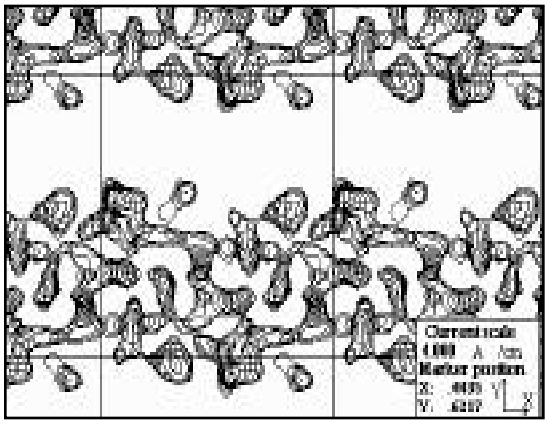
← α-helix



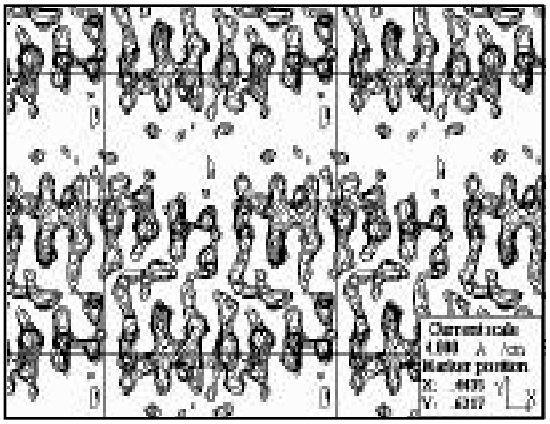
← β-sheet

Two regions of electron density are shown. The positions of the main chain atoms are shown (independent part only).

EFFICIENT RESOLUTION OF THE AB INITIO PHASED SYNTHESIS

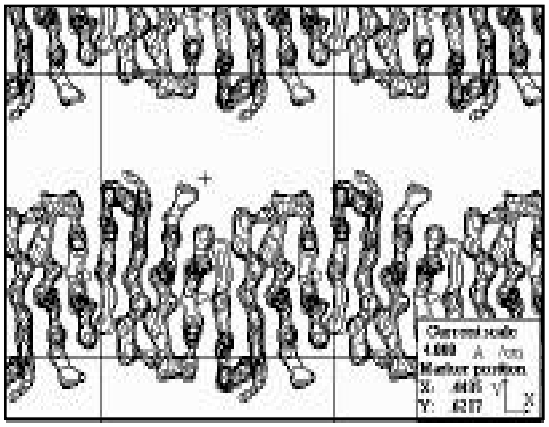


5Å resolution
experimental F^{obs}
exact phases



3Å resolution weighted synthesis

experimental F^{obs}
ab-initio phases
and weights



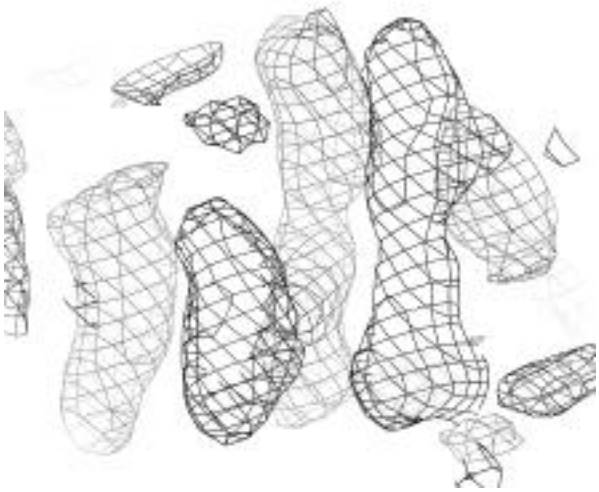
4Å resolution
experimental F^{obs}
exact phases

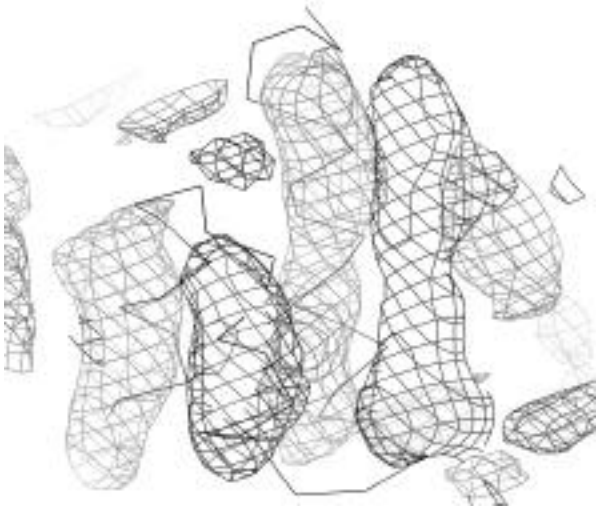
sections z= -6:6/42 are shown in projection along z-axes

DIRECT PHASING : ER-1

Anderson, Weiss & Eisenberg
(1996) *Acta Cryst*, D52, 469-480

Space group C2,
a=53.9, b=23.1, c=23.1 Å, β=10.4°

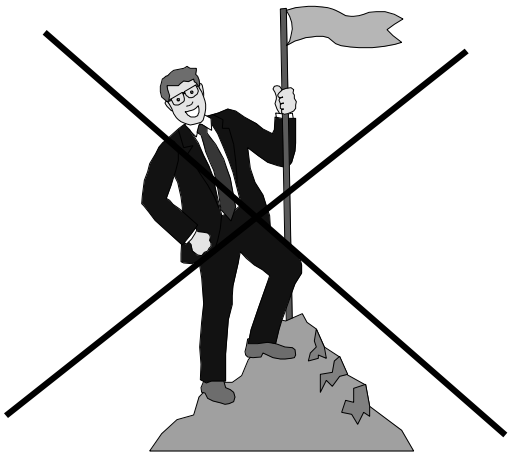




resolution	N° refl
∞-11	13
∞ -9	25
∞ -7	52
∞ -5	132
∞ -4	249

NEW SEARCH STRATEGY

There is no hope (or it is too weak ?) to identify unambiguously, among a number of generated variants, ‘a winner’, a set of phases close enough to the correct phase solution



Existing criteria allow to select, among the original ‘population’ of generated phase variants, another ‘population’ enriched by those close to the correct solution

A good approximation to the solution can be obtained by a simple averaging of the selected variants (eventually, a better result can be obtained using a clustering procedure)